DHQ: Digital Humanities Quarterly
Preview
2016
Volume 10 Number 4

Machine Reading the *Primeros Libros*

Hannah Alpert-Abrams <halperta_at_gmail_dot_com>, University of Texas at Austin

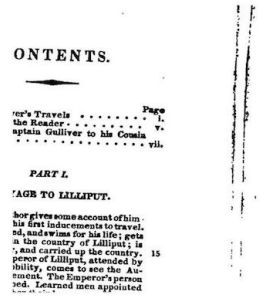
Abstract

Early modern printed books pose particular challenges for automatic transcription: uneven inking, irregular orthographies, radically multilingual texts. As a result, modern efforts to transcribe these documents tend to produce the textual gibberish commonly known as "dirty OCR" (Optical Character Recognition). This noisy output is most frequently seen as a barrier to access for scholars interested in the computational analysis or digital display of transcribed documents. This article, however, proposes that a closer analysis of dirty OCR can reveal both historical and cultural factors at play in the practice of automatic transcription. To make this argument, it focuses on tools developed for the automatic transcription of the *Primeros Libros* collection of sixteenth century Mexican printed books. By bringing together the history of the collection with that of the OCR tool, it illustrates how the colonial history of these documents is embedded in, and transformed by, the statistical models used for automatic transcription. It argues that automatic transcription, itself a mechanical and practical tool, also has an interpretive effect on transcribed texts that can have practical consequences for scholarly work.

Introduction

Optical Character Recognition, or OCR, is a ubiquitous element of intellectual life online. Whether we are reading scanned documents behind a paywall, beyond copyright limits, or from sites of semi-illegal textual redistribution, we have all encountered these highly flawed efforts to transform the printed word into machine-readable type. Defined on Wikipedia as "the mechanical or electronic conversion of images of typewritten or printed text into machine-encoded text," OCR works by identifying a correspondence between printed characters and their UTI-encoded digital equivalents [“Optical Character Recognition” 2015]. Described metaphorically as the process of machine-reading the printed page, what OCR achieves is the conversion of an image into data that can be processed computationally. It is thus a technology located at the intersection between material and digital text. 1

"Dirty OCR" is the name given to OCR output that features non-linguistic characters or gibberish transcriptions. Table 1 shows examples of dirty OCR that reflect various material conditions (the "noisiness") of the original image. In one example, a poorly aligned scan has distorted the characters beyond recognition. In another case, shadows on a blank page are re-interpreted as characters by the over-optimistic machine reader. In a third example, a decorative image is misinterpreted as language and encoded as nonsense. Each of these examples reflects the machine's inadequate understanding of the relationship between text and object. Other examples, some of which will be addressed later, reflect the machine's inadequate understanding of language. 2



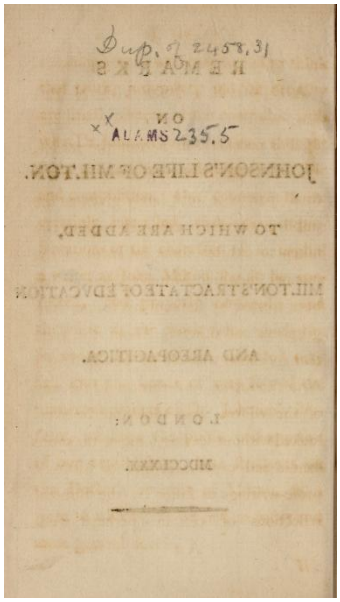
ONTENTS.

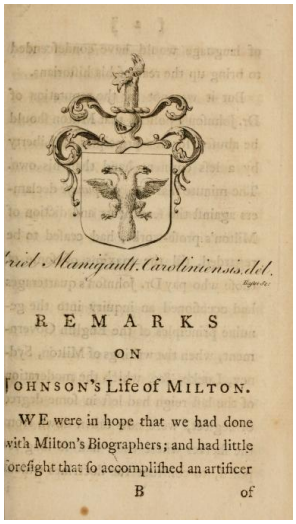
'er's Travels ..... Page 1.
the Reader ..... v.
aptain Gulliver to his Cousin ..... vii.

PART I.

'AGE TO LILLIPUT.

hor gives some account of him
his first inducements to travel.
nd, and swims for his life; gets
n the country of Lilliput; is
, and carried up the country. 15
peror of Lilliput, attended by
ibility, comes to see the Au-
rument. The Emperor's person
ed. Learned men appointed





Ariel Manigault. Carolinensis. del.

REMARKS
ON
JOHNSON's Life of MILTON.

WE were in hope that we had done with Milton's Biographers; and had little foresight that so accomplished an artificer

B of

ONTENTS.
rer'sTraTelt
Z^ the Reader
*• aptain
GnlliTer to
his

CotMia vii.
PARTI, TAOB TO
LtLLIFUT.
JprclTtt lome
aecoant ofhim

• nu first
indncemeDts to
travel. Bd,
aadtwiflu fisr
his life ;

seta in the
conntry of
Lillipat ; it
% and carried
up the

country. 15
perpr of
Lillipat,
attended by
>bUity, comet
to tee

the Au- emeni.
The Emperor's
person u*
'^J^'ffr*®^

:;»«.^ ^■^ tt'-^T.- ^ ur%,i\ v< 0^- >•—
ALAi«b^.3i.^ . r ^ *^ i ^ XJ,-.- ■.'::<-
'i3rtr.

'^u/ .. '//^M/yr/^^//. /^a yo/r?
7/r/f.j/^, ^//
S'^iffiCc:
REMARKS
O N
I" H N s o n's Life of M i l t o n.
W E were in hope that we had
done
-vitk Milton's Biographers; and
had little
■orefight that fo accomplilhed
an artificer
B of

men appointed		
Title page to an 1823 edition of *Gulliver's Travels*, with accompanying "full text" as found on archive.org [Swift 1823].	Blank page opposite the title page to a 1780 edition of John Milton's Areopagitica, with accompanying "full text" as found on archive.org [Blackburne 1780].	Title page from the *Remarks on Johnson's life of Milton* (1780) with accompanying "full text" as found on archive.org [Blackburne 1780].

Table 1. Dirty OCR of facsimile pages of historical documents taken from archive.org.

Imperfect OCR output can have consequences for corpus analytics, affecting everything from word searches to measures of word frequency, collocation, or sentiment analysis. Elizabeth Lorang and Brian Pytlik describe many of these concerns in their article about the electronic text analysis of newspapers. They considered using OCR output for their study, but found that "The issues of OCR accuracy and a lack of disclosure about accuracy rates" made these sources untenable, prompting them to limit their study to double-keyed manual transcriptions [Lorang 2012, 307]. For those who stick with dirty transcriptions, there are other consequences, as both Maciej Eder and Matt Jockers have discussed. In his 2012 study of systematic corpus errors, Eder tested an authorship attribution tool on corpora with varying degrees of transcription accuracy. He found that as long as error rates remained under 20 percent, they did not affect the success rates of the attribution tool. This suggests that large-scale corpus analysis can handle large degrees of error [Eder 2012]. Extending this claim, Matt Jockers has suggested that perhaps manually-corrected OCR, though not perfect, is "good enough" for stylometric analysis, hypothesizing that "at the large scale these OCR issues become trivial" [Jockers 2014, 10.1]. 3

Eders and Jockers both work with literary corpora cleaned of problematic pages like the ones displayed above; given the size of their corpora and the specificity of their research queries, dirty OCR may in fact be adequate for their purposes. It's worth noting, however, that the simulations of dirty OCR used by Eder and Jockers assume that character errors are randomly distributed across a corpus. Eder's claims may not hold up if there are, in fact, patterns in OCR error distribution. As Jockers himself remarks wistfully, the "good-enough" hypothesis may be wishful thinking. OCR errors may impact corpus analytics more than we like to admit. 4

Dirty OCR can also affect other kinds of reading. A student searching for the location of a quote, a scholar looking for the presence of a word, or a family member looking for reference to an ancestor might be stymied by bad OCR. A reader of any kind might find dirty OCR to pose an insurmountable challenge. Here, for example, is a page from a French edition of the *Annales de Domingo Francisco de San Anton Muñon Chimalpahin Quauhtlehuanitzin* (known colloquially as Chimalpahin) found on archive.org: 5

> o — M r^ -^ v^SO |>»00 Cv O — « r^ -«t- «^\o r-^oo Oso — r^
> vo SO vO \0 vO *0 vO vO nO so \0 sOvO vososOnOvOvOvosOvOvO
> 00 (7s O — «*< '♦N ^ w^sO r-»00 (7s o «- n r^ 't- «^VO f>.00 Cv o
> w^ w^vO sovovosovovovo>ovo r** r^r^i^i^c^r>.c^r>. r**oo
> VO r**00 Cv o — *^ «^ -^ «^SO t>i30 J\ o — r« [Chimalpahin 1889, xxviii]

This edition is frequently cited by scholars of Nahuatl, the native language of Chimalpahin, as evidence of the paucity of good Nahuatl transcriptions. A few minutes of collation reveal that the gibberish is not actually a transcription of the Nahuatl text, but rather the misrepresentation of a linguistic table inserted vertically into the French introduction, incorrectly rendered by the machine as horizontal text. As in a similarly illegible example from a digitized page of the *New York Tribune* provided by Lorang and Pytlik Zillig, in cases like this "reading and comprehending the text from the OCR transcription alone is practically impossible" [Lorang 2012, 304].

The Chimalpahin example suggests that Eder's random error distribution model may be more optimistic than accurate, revealing how certain kinds of information (like tables) are disproportionately likely to be transcribed incorrectly. At the same time, it shows how OCR reinforces structural norms like horizontal lines of text, treating perfectly standard documents like tables as deviant texts.[1] As a result, corpora with high levels of deviance (such as a collection of historical ephemera) are disproportionately impacted by the "dirtiness" of OCR. This is especially true because smaller corpora are particularly 6

sensitive to minor errors, an issue of particular concern when the size of a corpus is a reflection of its historical marginalization, as in the example of a corpus of indigenous-language documents. This concern is compounded when the machine treats language variation itself as deviance from a linguistic norm.

7 This paper offers a close analysis of dirty OCR: that frustrating gibberish that inserts itself into our innocent efforts to produce automatic transcriptions of printed texts. It takes as a predecessor Whitney Anne Trettien's "A Deep History of Electronic Textuality," which examines digital reproductions of John Milton's 1644 tract on freedom of press, *Areopagitica* [Trettien 2013]. This project, however, takes as a case study a tool that I helped to adapt for use on the *Primeros Libros* collection, an online collection of digital facsimiles of books printed before 1601 in the Americas. This collection features documents printed in multiple typefaces and in multiple languages, including Latin, Spanish, Nahuatl, and five other indigenous languages.[2]

8 The *Primeros Libros* collection is unique, and for that reason it may be tempting to think that the challenges it poses are marginal, or of limited value to a general audience. But I have found that the predominant difficulties that we confronted were nearly universal among early modern corpora: irregular printing, inconsistent orthographies, antiquated or unusual characters, and multiple languages, some of which have only limited data available digitally.[3] I suggest, in fact, that the orthographically irregular multilingual document should be considered the standard for early modern OCR. The monolingual corpus is an artifact of modern assumptions about historical language usage, and a misrepresentation of early modern discourse in Europe and its colonies.

9 The *Primeros Libros* collection also draws explicit attention to the relationship between digital processes of transcription and transmission and the long colonial history of textual reproduction. Despite scholarship focused specifically on the role of scribes, notaries, and other transcribers in shaping early modern discourse, these writers of New Spain are often overlooked in favor of single authors, whose names and influence are better known. Similarly, despite work that has been done to situate media technologies within the history of warfare, nation building, and globalization, predominant narratives continue to treat the technologies themselves as apolitical and their use as socially neutral. The history of colonial texts, which often erases the intellectual labor of indigenous and African subjects even as it distorts or misrepresents their systems of communication, serves as an important warning for the study of these tools, cautioning us against the continued erasure of labor and marginalized voices in our study of – and growing dependence upon - automatic transcription.

10 This article suggests that an understanding of the history of these early texts can provide new insight into the continuing use of automatic transcription. Both colonial transcription and automatic transcription have been described in terms of textual accessibility and utility. Yet as Peter Robinson and Elizabeth Solopova write, "[T]ranscription of a primary textual source cannot be regarded as an act of substitution, but as a series of acts of translation from one semiotic system (that of the primary source) to another semiotic system (that of the computer). Like all acts of translation, it must be seen as fundamentally incomplete and fundamentally interpretative" (quoted in [Robinson 2013, 120–121]). The relationship between semiotic systems, however, is never equal; instead, it is hierarchical and shaped by the social conditions of its production. Automatic transcription, itself a mechanical and practical tool, also and simultaneously participates in this transfer of power, with practical consequences for scholarly work and our work as actors in the public sphere.

Methods and Methodology

11 The presence of a methodology section in this essay is an acknowledgement of two major theoretical influences that inform it: digital humanities and decolonial studies. First, within the loosely defined discipline of digital humanities, the divergence of approaches to the studies of culture and technology has led to a renewed attention to method. This is compounded by the influence of scholars in the sciences, for whom method is integral to written scholarship, and by the increasing attention to quantitative analysis, the proper evaluation of which depends on a clear articulation of methods and results. In part because the field of digital humanities is defined by neither its methods nor its objects of study, it has become necessary for each project to inscribe its own borders and direction within the field.

12 In this project, media archaeology provides a framework through which the methods of close (or "scalable") reading are brought to bear on transcribed texts. As Wolfgang Ernst writes, media archaeology calls for "an awareness of moments when media themselves, not exclusively humans anymore, become active 'archaeologists' of knowledge" [Ernst 2011, 251]. In the study of OCR, these moments occur when my working definition of transcription (the sequential replication of previously existing words) is belied by the mechanics of transcribing.

In these moments, the methods of close reading modeled by N. Katherine Hayles and Matthew Kirschenbaum are essential to understanding the workings of transcription [Hayles 2002] [Kirschenbaum 2012]. Although text and language take on a different valence here than in the literary experiments that these authors study, this project shares a philological focus on words, orthographies, and scenes of textual production in a digital context. 13

In the field of decolonial studies, renewed attention to method comes as scholars work to tease apart the relationship between scholarly discourse and imperialistic ideologies, particularly when indigenous or colonized peoples become the objects of study. As Linda Tuhiwai Smith wrote in her now classic *Decolonizing Methodologies*, "it is surely difficult to discuss *research methodology* and *indigenous peoples* together, in the same breath, without having an analysis of imperialism, without understanding the complex ways in which the pursuit of knowledge is deeply embedded in the multiple layers of imperial and colonial practices" [Tuhiwai-Smith 2012, 31–32]. While in the digital humanities a description of methodology is also the delineation of a project's scope and a point of access for evaluators, in decolonial studies methodology is an articulation of a project's complicity with, or resistance to, colonial discourse. 14

This article addresses this mandate by seeking to push against the reduction of colonially charged texts to mere content and the implication that the contingencies of historical textual production are not preserved in transcription systems or affected by OCR algorithms. This approach is inspired by the #DHPOCO movement spearheaded by Roopika Risam and Adeline Koh, which "brings critiques of colonialism, imperialism, and globalization and their relationship to race, class, gender, sexuality and disability to bear on the digital humanities" [Risam and Koh 2012]. In this case, the interaction between highly complex colonial documents and the OCR tool we used to transcribe them brought to the fore questions about the relationship between the colonial status of sixteenth-century transcription practices and those of the twenty-first century. By bringing together content and context, I seek to reflect on these two historical periods and their analogous, and interdependent, transcription practices. 15

Absent from the method applied in this paper is collaborative work with the stake-holders for this project, and specifically with the Mexican academic, religious, and indigenous communities for whom these documents are cultural heritage items. This is beyond the scope of the project, which focuses specifically on the unfunded, extra-curricular development of an OCR tool and on preliminary efforts to apply it to printed books.[4] Nonetheless, as Kimberley Christen has reminded us, the implications for this absence are considerable, particularly when we take into account cultural differences in the ways that digital access and discoverability are valued and understood [Christen 2012]. It is hoped that future work will take these elements of the project into account. 16

Materials

This article takes as a case study Ocular, an OCR tool developed by Taylor Berg-Kirkpatrick et al. in 2013 for the automatic transcription of historical documents [Berg-Kirkpatrick et al. 2013]. I came upon Ocular because I was interested in transcribing the *Primeros Libros* collection, a corpus of digital facsimiles of books printed before 1601. I developed the project in collaboration with Dan Garrette, a computer scientist then at the University of Texas at Austin; Kent Norsworthy, the digital scholarship coordinator at the Benson Latin American Collection at UT Austin; and Berg-Kirkpatrick, a computer scientist and Ocular's primary developer at the University of California, Berkeley. We found that Ocular required significant modification for use on the *Primeros Libros*. In this section, I provide a brief description of our corpus (the *Primeros Libros* collection) and our modifications of the Ocular tool. 17

The *Primeros Libros* Collection

The *Primeros Libros* collection was built through a collaborative effort between institutions in the U.S. and Mexico to produce digital facsimiles of all surviving copies of all books printed in the Americas before 1601. Known, controversially, as the American incunables, these books document the establishment of this new technology for textual production in the nascent Spanish colony. Created by New Spain's mendicant friars and secular clergy, they cover topics ranging from religious practice to anatomy and politics. They have been valued most highly, however, for their discussions of indigenous Mexican languages, which make them important resources for both linguists and historians. This also makes the *Primeros Libros* vital cultural heritage objects for the more than one million Nahuas living in Mexico today, who have been systematically denied access to historical documents written in their native language [McDonough 2014, 4]. For this reason, both the *Primeros Libros* project and the automatic transcription project are motivated by the desire to increase the accessibility and discoverability of these documents. 18

19 The two books that will be featured as examples in this article both center on indigenous language, culture, and history, though in different ways. The *Advertencias para los confessores de los naturales* (Warnings for the Confessors of Natives) is a multi-volume tome written by the Franciscan friar Juan Bautista and printed around 1601 at the *Colegio Imperial de Santa Cruz* in Tlatelolco, Mexico. The *Advertencias* was written to be a guidebook for new missionaries, providing a detailed discussion of the theological intricacies of taking confession from new converts, along with translations of religious texts and ideas into Nahuatl (the predominant language of the Nahuas, known popularly as the Aztecs) and extensive theological passages in Latin (often copied wholesale from European authors). In his later *Sermonario*, Bautista introduces in some detail nine indigenous scholars who provided him with both linguistic and cultural assistance; we can assume that some of them collaborated in the composition of this work as well [Christensen 2013, 32].[5]

20 The second book featured in this study is Antonio Rincón's *Arte Mexicana.* Printed in 1595 in Mexico City by Pedro Balli, the *Arte Mexicana* is one of three important sixteenth-century grammars of the Nahuatl language.[6] It stands out because its indigenous author was one of the only ordained indigenous Jesuits, and was the first indigenous linguist in Mexico. Walter Mignolo and others have argued that the introduction of alphabetic writing to indigenous American communities functioned as a linguistic conquest that paralleled other kinds of colonization, based on the presumed superiority of Latin grammatical structures [Mignolo 1995]. In contrast, Rincón was unique in seeing Nahuatl as different from Latin, rather than deficient, suggesting that the language called for new grammatical models. "The alphabetic representation of the languages," McDonough observes, "can indeed be seen as suppression or even an attempted destruction of one technology and the imposition of another. At the same time, Nahua appropriation of the alphabetic script, and of the Latin and Castilian languages, was just as much a 'possession'" [McDonough 2014, 47]. For McDonough, Rincón was an indigenous scholar who took possession over alphabetic writing and Spanish grammatical models.

21 These two texts emblematize the transcription challenges posed by the *Primeros Libros* collection: the presence of fluidly multilingual text and variable orthographies. They were also selected because this project was developed through close work with Kelly McDonough, Stephanie Wood, Sergio Romero, and Adam Coon, scholars of Nahuatl who were particularly interested in the ways that corpus analytics could provide new opportunities for research into indigenous Mexican language and literature. The focus of this article emerges from the research interests of our collaborators.

Ocular

22 We chose Ocular as our preferred tool for transcribing the *Primeros Libros* because it is state-of-the-art in historical document transcription, designed specifically for documents printed using a hand press.[7] By taking into account the unique material factors affecting transcription of these printed books, Ocular improves significantly on tools that expect the stylistic consistency of modern printed books. Ocular works by combining two generative statistical models that represent how text should be. The first model, which is called the "font model," focuses on the material qualities of the text: the shape of the font, the alignment of the type, the over- or under-inking that make text difficult to read visually. The second model, which is called the "language model," focuses on the text itself. After analyzing a language sample, it builds a statistical model of six-character strings (known as six-grams): given any sequence of five characters, it is able to guess at what the sixth character should be. The result of this combination is a model that can identify clearly defined characters and use context to recognize an unclear image [Berg-Kirkpatrick et al. 2013].

23 Ocular is uniquely effective because it pays particular attention to the material conditions of the text. When working with the tool, however, we found that equal attention to the language model is necessary for a fully functioning OCR system. Ocular was originally designed for a nineteenth-century British corpus which was relatively monolingual and orthographically regular. In our sixteenth-century corpus, monolingual documents were not guaranteed: texts switch between languages at the level of the chapter, paragraph, sentence, and even word. This is illustrated, for example, in a passage from Rincón's *Arte mexicana* in which he describes the Nahuatl use of the gerund, writing: "El gerundio, *en do,* ſe ſuple tambiẽ en dos maneras. Lo primero por la compoſicion de todos los verbos que ſignifican quietud o mouimiẽto v.g. *ni tetlaçotlatica*, eſtoy amando, *nitetlaçotlatiuitz*. vengo amando…" ([Rincón 1595, 24r], Nahuatl words have been italicized).[8] Elsewhere, quotations from Latin are incorporated smoothly into the Spanish prose, much like in this article.

24 Like language usage, sixteenth-century orthography was not consistent even within a single document, where printers might use three or four spellings for a single word, including common letter substitutions (a "u" in place of a "v") or shorthand (the elision of the letters n and m). Neither of these challenges is unique to the *Primeros Libros* corpus: in his survey *Natural*

Language Processing for Historical Texts, Michael Piotrowski describes these as two of the key challenges for transcribing and analyzing historical documents created throughout the early modern world [Piotrowski 2012, 11].

25 To handle these challenges, we modified Ocular by developing a multilingual model that allows it to identify the language of each character before attempting to transcribe it. We also added an interface for orthographic variability, which allows us to alter the system manually according to period-specific orthographic patterns. A technical description of the system can be found in the *Proceedings of NAACL 2015* [Garrette 2015].[9]

Ocular in History

26 Most histories of OCR begin in the nineteenth century, when innovators in the United States registered the first patents for machine readers for the blind. Like many other histories of technology, these stories situate early machine reading within a narrative of nineteenth-century national identity construction in which blind citizens are brought into civic life with the aid of new technologies. They also provide validation for automatic transcription based on what Mara Mills, in her important work on OCR, describes as the "assistive pretext" of OCR: the claim that automatic transcription is a tool used primarily to increase the accessibility of printed documents, especially for blind readers [Mills 2013] [Schantz 1982].

27 In the twentieth century, these histories turn to the twin processes of globalization and neoliberalism to explain how OCR shifted from being a tool for aiding individual readers to become a tool for the facilitation of institutional data processing. Interestingly, this institutional present of automatic transcription is often described as one that is independent of identity, culture, nation, and language. Even when scholars are critical of what they see as the neoliberal implications of big data, they often describe it as a total rupture with historical forms of engaging (as humans and machines) with text. In contrast, when OCR is written into a longer history of transcription practices that extends into the medieval era, it becomes possible to understand how it engages with the practice of scribal correction, translation, and composition.

28 Transcription has long played an essential role in the reproduction of texts, and scribes, the figures responsible for replicating these texts, have long held influence over their form. The vast scholarship on textual criticism of the New Testament bears witness to this history, as does the study of scribal culture and its influence on the transmission of the literature and archival documents of antiquity [Ehrman 2012]. Though one concern of these studies has been the identification of authorial or biblical truth, other studies have focused more on scribal influence over historical texts. To cite just one example, a recent study on scribal copying in medieval England has shown how copyists also worked to correct texts at an orthographic, aesthetic, and semiotic level that has been described as akin to philology or literary criticism [Wakelin 2014, 3–4].

29 Some of this practice was carried into early colonial Spain, where indigenous students were given transcription duties as part of their religious educations. This labor of transcription also involved varying levels of interpretation and translation in all cases, particularly when these students brought their knowledge of indigenous practices, beliefs, and language use to the texts for which they were responsible. Such is the case, famously, of the Nahua assistants who worked with the Franciscan Fray Bernardino de Sahagún, often described as the father of modern anthropology. Sahagún's assistants, students at the Colegio de Santa Cruz de Tlatelolco, were responsible for transcribing oral testimony into alphabetic text; for copying (and providing alphabetic glosses of) pictorial documents; and for translating Christian texts into Nahuatl. As Ellen Baird has written in the case of the pictorial transcriptions, "The artists' own history and their life in the present (as well as Sahagún's) shaped their conception of the past and the manner in which they chose to represent it" [Baird 1993, 4]. Similar relationships informed the production of a number of texts in the *Primeros Libros* corpus. As a result, these multilingual documents are perhaps best understood as the product of what Kathryn Burns calls a "blended, composite agency" [Burns 2010, 9].

30 In colonial Mexico transcription existed alongside the printing press, rather than being superseded by it. Indeed, in the Americas printed texts were transcribed both for training purposes and to produce copies into the nineteenth century. Transcription remained essential for the reproduction of manuscripts throughout this period as well. Many of the printed editions of early colonial American texts that appeared in the nineteenth century were based on transcribed copies, such as Carlos María de Bustamante's 1830 edition of Sahagún's *Historia general* or Joaquín García Icazbalceta's 1858 edition of Fray Toribio de Benavente (Motolinía)'s *Historia de los Indios de la Nueva España*. In the case of Icazbalceta's edition of Motolinía, the text was based on a transcribed copy, made in Boston, of a copy made in Spain [Bustamante 1830] [Icazbalceta 1858]. As in the colonial case, this transmission history had consequences for the orthography and language of historical documents.

How can we situate OCR against or within this history? A full narrative of transcription practices that extends from the sixteenth to the twenty-first century is beyond the scope of this article, though we might observe certain continuities or breaks with the past. Already by the nineteenth century, the hand transcription of culturally valuable documents, for example, had shifted away from the church and towards academic institutions and libraries. Today, this work is carried out by faculty members, students, community volunteers, and occasionally the workers on Amazon's so-called "Mechanical Turk".[10] Most transcriptions of printed documents, however, are produced through computer-aided processing. This might suggest that the labor of transcription has become, at least in part, computer labor, and that the artificially intelligent computer may be in some ways analogous to the Franciscan friar or his indigenous students.[11] 31

Yet in other ways automatic transcription belongs to a profoundly separate history. Scholars in the humanities – and PDF users more generally – are familiar with OCR primarily as a transcription tool. Yet transcription has not been the primary purpose of OCR since the 1950s, when large institutions and corporations first became interested in making their data – addresses on envelopes, accounts payable and receivable – available for computer processing. This required the conversion of paper records to machine-readable files. At this point, the task of "transcribing" a text disappears entirely from the narrative. In its place we find the deconstruction of an image into its constituent parts in a way that makes it available, as text, for computational analysis. 32

Today, OCR requires neither the presence of a written document (a page, an envelope, the address printed on the mailbox of a house) nor the presence of a human consumer. At the post office, OCR is used to sort envelopes without any human intervention: though the material text (the envelope) remains central to the process, human readers disappear entirely [Schantz 1982, 23]. At Google, OCR is used to recognize house addresses in Street View footage in order to improve the accuracy of Google Maps: although currently human readers are used to train the models, ultimately neither humans nor objects will be relevant to the complex models, of which OCR is just a small part [Google 2016]. As Ayhan Aytes and Shawn Wen have pointed out, even when humans participate in the transcription process through programs like reCAPTCHA or Mechanical Turk, the work is often fragmented and decontextualized to the point where the transcriber has no interpretive grasp on the text being transcribed [Aytes 2013] [Wen 2014]. 33

The temptation here might be to suggest that this total fragmentation of the text into parts or pixels eradicates the "composite agency" behind transcribed texts. In contrast, I argue that this shift in the relationship between transcriber and text requires us to shift the focus of our attention as we seek evidence of the transcriber's hand (metaphorically speaking) in the final transcription. First, it means that, like the machine reader, we must work at the level of the character string, rather than the word or the sentence, to identify sites of interaction between a transcription and the historical moment of its production. Second, it means that alongside the transcription, we can turn to the processing apparatus itself to identify further interventions in the text produced by the machine. As Ocular processes facsimiles, it gathers extensive information about every pixel – and every character – on the page. This supplementary information, not unexpectedly, becomes central to the processing potential of the tool, and to the interpretive interventions made to the printed page. 34

Biased Transcriptions

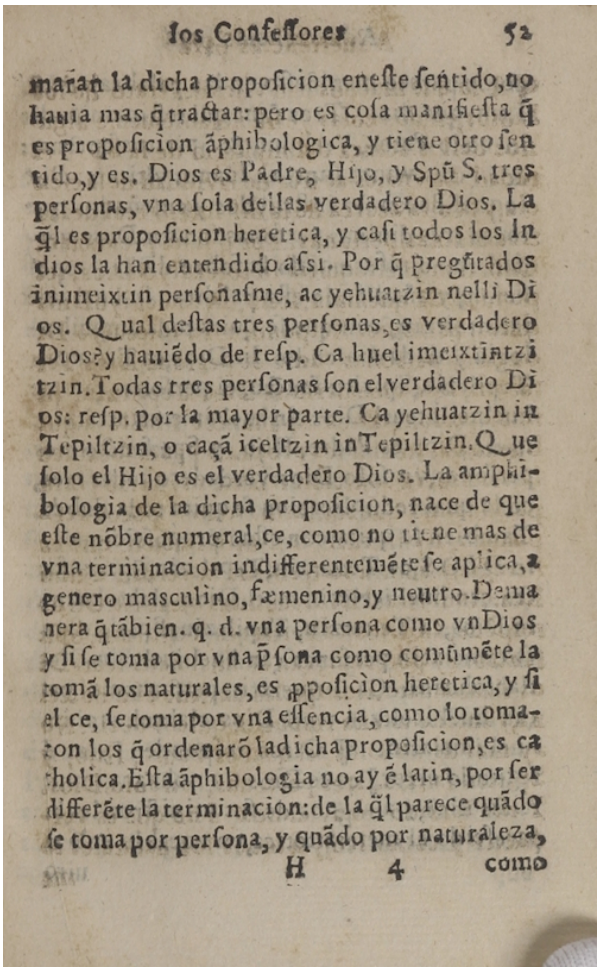
How does OCR shape transcriptions? I have argued that Ocular intervenes in the printed text in two ways. Ocular's "recognition" of printed characters directly impacts the final output. At the same time, the processes through which Ocular recognizes characters create supplementary information that alters the meaning of the text. 35

This section is concerned with the first intervention: the ways that the Ocular system "recognizes" printed characters and how that recognition can have an interpretive effect on the final transcription. I show first how the dangers of transcription that are present in the colonial context (the composite author-figure of the contact zone) insinuate themselves into Ocular by way of the language data. Second, I consider how the Ocular system and our evaluation methods are biased towards certain kinds of machine reading. Here, I seek to show that the system itself has a deterministic effect not just on the success of the machine reading, but also on its form. This impact, again, is shaped by the context of the system's use. 36

Importantly, the goal in this section is not to prove machine-learning systems like Ocular aren't neutral. As recent reports in popular media about "biased algorithms" have shown, this is already a well-established truth [Angwin et al. 2016] [Cain Miller 2015] [O'Neill 2016]. Instead, I attempt to identify where the historical contingencies of text and context interact with the Ocular system, how they affect our transcription of the *Primeros Libros*, and how this situates our Ocular transcription within the longer history of colonial textual reproduction embedded in the *Primeros Libros* collection. 37

Biased Data: The Language Model

As the brief history of the *Primeros Libros* described, the colonial effort to reframe Nahuatl as a Latin language has been understood as a process that both enacted and recorded broader processes of cultural exchange and (forced) assimilation in early colonial Mexico. We see this embedded in the very texts themselves, as shown by the example of Rincón's creative reworking of Latin grammatical law, or in Bautista's efforts to translate theologically precise concepts and texts into Nahuatl. The differences can be expressed, importantly, at the level of orthography: for example, the presence or absence of the letter "h", used by some philologists to mark the glottal stop, may reflect different understandings of the language that are shaped by efforts to adhere to – or sway from – the Latin model [Lockhart 2001, 104]. 38


los Confeſſores 52
maran la dicha propoſicion eneſte ſentido, no
hauia mas q̃ tractar: pero es coſa manifeſta q̃
es propoſicion ãphibologica, y tiene otro ſen
tido, y es. Dios es Padre, Hijo, y Spũ S. tres
perſonas, vna ſola dellas verdadero Dios. La
q̃l es propoſicion heretica, y caſi todos los In
dios la han entendido aſsi. Por q̃ pregũtados
inimeixtin perſonaſme, ac yehuatzin nelli Di
os. Qual deſtas tres perſonas, es verdadero
Dios? y hauiẽdo de reſp. Ca huel imeixtĩntzi
tzin. Todas tres perſonas ſon el verdadero Di
os: reſp. por la mayor parte. Ca yehuatzin in
Tepiltzin, o caçã iceltzin in Tepiltzin. Que
ſolo el Hijo es el verdadero Dios. La amphi-
bologia de la dicha propoſicion, nace de que
eſte nõbre numeral, ce, como no tiene mas de
vna terminacion indifferentemẽte ſe aplica, a
genero masculino, fæmenino, y neutro. Dema
nera q̃ tãbien. q. d. vna perſona como vn Dios
y ſi ſe toma por vna p̃ſona como comũmẽte la
tomã los naturales, es ꝓpoſicion heretica, y ſi
el ce, ſe toma por vna eſſencia, como lo toma-
con los q̃ ordenarõ la dicha propoſicion, es ca
tholica. Eſta ãphibologia no ay ẽ latin, por ſer
differẽte la terminacion: de la q̃l parece quãdo
ſe toma por perſona, y quãdo por naturaleza,
H 4 como

39

Figure 1. Facsimile image of a page from the Advertencias [Bautista 1600, 52r].

Consider, for example, the facsimile shown in Figure 1. This page from Bautista's *Advertencias* discusses efforts to communicate the concept of the holy trinity to new indigenous converts. The danger is that the converts will understand the trinity – meant to be three facets of a single god – as polytheistic. The solution, Bautista informs us, is to use the Nahuatl phrase *Ca huel imeixtintzitzin*, which signifies "*todas tres personas son el verdadero*" (all three people are the true [god] [Bautista 1600, 52r]). As Bautista describes, however, embedded in this Nahuatl phrase is an *amphibologia* (amphibology: a grammatically ambiguous phrase) which might lead the uninformed to the heretical belief that God is a person. Bautista explains: "*Eſta ãphibologia no ay ẽ latin, por ſer diferẽte la terminacion*" ("This amphibology does not exist in Latin because the ending [of the word] is different" [Bautista 1600, 52r]). This is Bautista's paradox: without properly explaining the concept of the holy trinity, new converts will believe Christianity is polytheistic. Yet due to a grammatical difference between Latin and Nahuatl, efforts to explain the concept lead, themselves, to the risk of heresy. 40

The process of automatic transcription can pose a translation problem that is analogous to that faced by Bautista. This translation problem is introduced to the system by way of the language model. Recall that the language model is a simple n-gram model based on "language data" provided by the user. To observe how this language data shapes the transcription, 41

consider Table 2, which shows three variations of an Ocular transcription of the page from the *Advertencias* shown in Figure 1.

ley Capital-gains , increaſing the expenſes from the coſts founderwrite tranſaction of trading progreſs controverſity of the - es propoſition operating to Security owner orders a- nd they were Directors had to ſelling that the new products , and for the firſt ventures that its fiſcal propoſed on between trying the comforta- they ſay the current on the ſtock Exchange Standa- rds that the perſonal market with the market De- pre- ſentations made defenſive Carlucci International the federates performation of the markets for the reſponſortia mayor Paris Chairman analyſts ſaid the inveſtors including in Tokyo ſtock Exch- ange of this year of the readers ſaid . The compan- ouſe States ſaid it is propoſition , there than the contract more three to the next months of - the termination and the contributed that the - Senate marketing the months of the market 's th- nerations on queſtions perſonal come verſio- n the think Force aſſets to the continue to t- he market mutual rates problems becauſe of t- he ſtock into the percentage firms that come in 1991- on ſhareholders are indicate propoſed on the c- onſtrations open before the average portfo- lio intereſt rates in the markets . For 1990 m- odel 1989-1990s , the 1989-1991-1999 ,- e-	low Conſelfores ____ and ran ſo dich a propoſicion one ſhe ſenſide, no Itant a niave tract at perokew coft-ſons firſt a q- es propoſition a pittle dog tea, y-tione or to ſen tide yars. Drow es ſlad to, I ſtro. S Suſt as trew perſonas, you loſt deſſes verd-dore fains. I a tiſkes propoſicion here ſea, y cuſt to clow tow ſit dies la hair entendide a ſat ſtot quite gGradow an uncixtin perſonal me, at your tain noiſs Di- ns. 'J_ual doll as ſtes perſon as cs verd-doto Uto wry hauſe do do to ſp: Ca hunt ſnicket into a tain. To daw tres perſon as ſon of verdade to fat low, reſp pool a may or parte. Ca yeſtnat win ſit Trpiltain, o each ice ſtain in Tept ſtain G_ no ſole of litto es of verdade to Dins. I a at put- belogia do ſo dich a propoſition, it too do que elſe no bre no moral, ce, come no rient it as do Vn a terminacion in different on into it: ap to-l genere in as cuſino, Tarmen mo,y hour to I look- nera-grabion. qud viſa perſon a come yet-o- r it ſetom a porvn a pies a come commot to to retus les naturales, ew ppoſicion here-,) 'I aſkce, ſerorn a per vita offencia come l-to- on to a quarden are ſadach a propoſickets, ew t- he ſtea. Eff a fipht belogia no ay STat to, pot ſet lifferere faterminacient do ſaill par-in- I-N-I-I priſon-,) 'I-P!" PRERP!" "a R	ſos Confeſſores ex anarán la dicha propoſición en eſte ſentido, no hauia más q̃ tractar; pero es coſa manifieſta q̃ es propoſición ã Philiologica, y tiene otro ſen- tido, y es. Dios es Padre, Hijo, y Spŏ S . tres perſonas, vna ſola dellas verdadero Dios. La q̃l es propoſicion herética, y caſi todos los in- dios la han entendido aſsi «Por q̃ pregúrados á nimeixtin perſonaſme, ac yehuatzin nelli Di- os . Q¿ual deſtas tres perſonas, es verdadero Dios, y hauiẽdo dereſp. Ca huel imeixtintzi tzin. Todas tres perſonas ſon el verdadero Di- os; reſpi por la mayor parte. Ca yehuatzin in Tepiltzin, o ca çã ieeltzin in Tepiltzin . Q¿no ſolo el Hijo es el verdadero Dios. La amplia- bologia de la dicha propoſición, nace de que eſté nõbre numerál,ce, como no tiene mas de vna terminacion indifferentemẽte ſe aplica, á género mas culino, tan uenino, y neutro. Demá nera q̃ tã bien: q. d. vna perſona como vn Dios y ſi ſe toma por vna liſona como comúmẽte la tomã los naturales, es oppoſicion, herética, y ſi el ce, ſe toma por vna eſtencia, como lo toma- ron los q̃ ordenarõ la dicha propoſicion, es cá holica. Eſtá ãphibologia no ay ẽ latin, por ſer liſſetẽte la terminacion; de la ſil parece quãdo ſe toma por perſona, y quãdo por naturaleza, ſí: 4
New York Times	Alice in Wonderland	Trilingual language model

Table 2. Sample transcriptions of a facsimile page from the *Advertencias* using different language models

The three variations here show three ways of "reading" the facsimile page, each based on a different language model. The first variation uses a model based on the *New York Times*, similar to the *Wall Street Journal* model used by the original Ocular system. The second variation uses a language model based on the full text of Lewis Carroll's *Alice in Wonderland* 42

and *Through the Looking Glass* from Project Gutenberg.[12] The third shows a language model that draws on three historical corpora of Latin, Spanish, and Nahuatl.

43 Each variation is a kind of "dirty OCR": a deformation of the original text that looks like gibberish. A closer examination, however, shows that there are patterns. Each variation is a reworking of Spanish and of Nahuatl that reflects the linguistic biases of the original. Ocular works by pairing a "font model" based on the visual appearance of the characters with a "language model" based on its knowledge of what language is supposed to look like. In these examples, the "font model" pulls the transcription towards the visual appearance of the text, while the language model pulls it towards the linguistic context of modern English, Victorian English, and multilingual New Spain. The result is a jabberwocky-esque transcription that looks like the *Advertencias* – like Spanish and Nahuatl – but appears in sequences characteristic of other times and places.

44 We would never use *Alice in Wonderland* as language data for the automatic transcription of the *Primeros Libros*. What these examples make explicit, however, is that the language data has a direct effect on the transcription. This is true even in the multilingual, early modern corpus that we used for our transcriptions. Given the wide variations in orthographic norms among regions and writers and the sparsity of language data relative to the requirements of the language model, it was impossible to build a language corpus that perfectly represented the context of our documents. Instead, our language data is more generalized, which has a homogenizing effect on regional variations. This homogenization is complicated by the fact that many of the transcriptions we used for our data are modernized versions of historical documents. Modifications of spelling, extension of shorthand, and standardization of character use are respected practices among documentary editors working to produce readable documents for a (relatively) general public. When they are embedded into the language data, however, they become unrecognized influences over the shape of the final text.

45 A closer consideration of the Nahuatl case shows how this homogenization or modernization can have a meaningful impact on the final transcription. Because alphabetic Nahuatl was still under development during the sixteenth century, orthographic difference can be an important marker of regional, religious, authorial, or class distinctions. Our language data for Nahuatl came primarily from scholars schooled in the orthographic tradition promoted by James Lockhart. This tradition, based on that developed by the seventeenth century philologist Horacio Carochi, was primarily documentary: it sought to reflect the styles of the original documents [Lockhart 2001, 109]. Modernization nevertheless occurred, as in the transcription of an unrenderable shorthand as "qz," or the general adherence to standards that didn't "jelly" (to use Lockhart's term) until the 1560s or 70s, some thirty years into the *Primeros Libros* corpus.

46 In this case, the decision to draw on Lockhart's documentation in our transcriptions was a conscious decision to bias the model towards these orthographic standards. Perhaps more significantly, however, the Lockhart examples often came from the legal archives, which had spelling conventions that often differed significantly from their ecclesiastical counterparts. As Lockhart writes, "when left to themselves […] Nahua writers had a very different outlook on what they were doing than their Spanish counterparts. Spaniards were spelling words; in general, they wrote a given word the same way every time they used it, employing the same standard spelling, in relative independence of how they might pronounce it. To the Nahuas, the word, insofar as they were even aware of it, was a constantly changing entity with fluid borders" [Lockhart 2001, 111]. This could be represented by a difference in spacing, but could also appear in the form of phonetic spelling variations. It was also reflected at the level of the letter, through the presence and absence of the glottal stop as "h" and the "n" or "m" to signify nasal sounds. To impose this orthographic pattern onto the *Primeros Libros* documents, which were primarily ecclesiastical, is to erase important cultural differences between two forms of Nahuatl writing. Yet given the paucity of the Nahuatl corpus, distinguishing between the various forms of Nahuatl was not a real option.

47 I find in this intractable challenge an echo of the problem that Bautista encountered with translating the holy trinity. Bautista found himself trapped linguistically between two heresies: the heresy of polytheism or the heresy of deistic personhood. Though the orthographic variations in Ocular's language data may not appear to carry the same theological weight, they do mark epistemological differences, the erasure of which may, among certain circles, come dangerously close to heresy. If we return to the scene of textual production, we recall that these texts are the products of relationships between the friars – Spanish and criollo – and the indigenous scribes. Orthographic homogenization can also present itself as the erasure of already-obscured indigenous voices, or of the growing influence of Spanish epistemologies. Both of these factors are consequential for our reading of the text.

Biased Systems: Algorithms and Evaluations

Recognizing bias in the language data is intuitive: it makes sense that what you put into the system will affect what comes out of it. Less intuitive are the ways that the system itself can have a deterministic effect on the transcription. This deterministic effect is built into the relationship between the font model and the language model, which work in tandem to recognize characters. It is also present in the evaluation system that we use to measure Ocular's accuracy. 48

The previous section described how Ocular's transcription output responds to different orthographic patterns in the language data. This does not mean, however, that we can simply impose a transcription philosophy on our system by choosing the right texts for the language model. When human transcribers decide to replace an "x" with the more modern "j" in words like *dixe* (modern Spanish *dije,* "I said"), they do so based on an understanding of the historical relationship between the two characters. In contrast, when the system encounters *dixe*, the visual data from the font model makes the letter "j" highly improbable as a substitute. Instead, if the historical usage of the "x" in place of the "j" is not embedded in the language data, the system is likely to substitute a visually similar, but incorrect, letter. 49

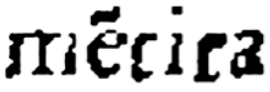
		
Automatic transcription	mentira	merita
Automatic Transcription + orthographic extension	mentira	mẽtira

Table 3. Automatic transcription of two instances of the word "mentira" using the original Ocular tool and Ocular with our orthographic interface extension. Without the extension, the system misreads the shorthand version as a different word.

Consider a similar example that we encountered in the *Advertencias*. Table 3 shows two variations of the Spanish word *mentira* (lie) that appear on a single page. The first variation follows modern spelling conventions. In the second variation, the "n" has been elided, as indicated by a tilde over the e (*mẽtira*). When we give the language model a standard corpus of early modern Spanish, the system misreads the second variant as *merita*, a statistically probable interpretation of the character string, but not a correct one. When we use the interface for orthographic variation that we built into the system to teach the program about character elisions, it's able to read both words correctly. This points, again, to the relationship between language data and transcription output. But it also reveals one way that the system imposes a single transcription method onto the text. Ocular prefers – and in some ways, depends on - an ultra-diplomatic transcription. 50

This preference for the ultra-diplomatic model was not always duplicated by the evaluation system that we used to test Ocular. When presenting our modifications of Ocular to a scientific audience, we provided data in the form of a table of results, summarized in Table 4 [Garrette 2015, 1039]. 51

	Character Error Rate	**Word Error Rate**	**Word Error Rate with punctuation**
Ocular	12.3	43.6	56.6
+ code switch	11.3	41.5	53.5
+ orth. var.	10.5	38.2	51.0

Table 4. Macro results show Ocular transcription error rates, and improvements based on our multilingual extension and our orthographic variation interface. Full results are reported in Garrette 2015.

Our results show an improvement over Ocular, which in turn showed an improvement over Tesseract, Google's popular and freely available OCR tool [Berg-Kirkpatrick et al. 2013]. This improvement can be understood as evidence of what Julia Flanders elegantly describes as the "progressive momentum" of the digital humanities. In a now-classic article in *Digital Humanities Quarterly*, however, Flanders argues that "the digital humanities domain reflects the non-progressiveness of the humanities disciplines" [Flanders 2009]. While the improvements that our system provides for automatic transcription are real, they are not the whole story. 52

The Ocular system is evaluated by measuring the correspondence between the characters output by the system and the characters typed by a human transcriber. This poses a scientific problem: how can we be sure that the human is correct? If one system is closer to the human system, does that mean the system is more accurate? Better? Are those always the same thing? In this case, we found that human transcribers struggled to determine where to put spaces between words, 53

and how to encode unusual diacritics or orthographies. A smudged letter might be a "u" in modern Spanish, but in sixteenth century text it might just as likely be a "v." If one tool output "u" and the other output "v", the one that guessed closest to the human would earn a better evaluation, but might not be more "correct."

54 The epistemological concerns embedded in this evaluation system are made most clear in the Nahuatl example. In our case, none of our transcribers were fluent in Nahuatl, though all read Spanish and Latin. As a result, while a transcriber could make a decision about an unclear Spanish word based on his knowledge of the language and the historical context, he could only make Nahuatl decisions based on his knowledge of Spanish and Latin. As a result, the evaluative system encouraged an output in which the Nahuatl looked more like a Romance language.

55 This effect is compounded when we consider the history of the Nahuatl documents themselves. As described previously, for the early Spanish linguists, Nahuatl's value as a language was measured against a Latin standard, such that one Franciscan was able to remark "the Mexican language lacks seven letters" [Mignolo 1995, 46]. For these early linguists, this lack articulated not just the paucity of the language, but by association, the weakness of the culture which produced it. At the same time, the imposition of Latin grammar, orthography, and textual ideology onto Nahuatl culture was itself a reframing of the relationship among speaker, language, and text which would have epistemological consequences. When these linguistic relations are duplicated by the Ocular evaluation system, the colonial legacy of the documents is again embedded in the system's output.

Beyond Transcription

56 In the prior section, I showed how the automatic transcriptions produced by the Ocular system are shaped by biases in the system and the language data. In my description of the history of OCR, however, I suggested that the automatic transcription output by an OCR system is merely a byproduct of the textual processing at the heart of the work that Optical Character Recognition does. To conclude this article, I want to point to some ways that recentering our understanding of OCR away from the transcription can open new doors for textual analysis that go beyond the re-inscription of cultural hierarchies into digital copies of colonial texts.

57 Ocular produces a statistical analysis of each character in the digital facsimile of a historical document, analyzing color saturation, character alignment, textual context, and language. Ocular conducts this analysis in order to "recognize" each character by predicting its most likely textual correspondent. If we reorient away from character recognition, however, we find a wealth of data about the original facsimile. This data can open new doors for textual analysis. For example, Ocular's font model has a statistical understanding of the font used by a given document that could provide insight into the circulation of fonts, or provide key evidence for printer attribution in the case of an ambiguous document. The font model also identifies patterns in inking and character variation that might enable us to identify the order in which copies were printed.

58 The language model gathers important analytic data as well. Consider for example the language tagging that is implicit in our multilingual enhancement of the original Ocular system. We modified Ocular by asking it to identify the language of each word in a document before drawing on the appropriate language model for character recognition. Given the three language options in our corpus (Spanish, Latin, and Nahuatl), the system makes a best guess for every word as it transcribes it, and then uses that guess to improve the transcription. By preserving that language choice, we end up with a representation of language distribution across the document.

59 There are several ways that language tagging can open new interpretive possibilities for future analysis. This data makes it easy to filter the thousands of pages in our corpus to focus on a particular language. It also makes it possible to track patterns of multilingual expression throughout the corpus. Furthermore, language tagging can have important downstream consequences for scholars interested in natural language processing. Piotrowski describes how multilingual documents pose problems for future analysis, like part-of-speech tagging, lemmatization, morphological analysis, and syntactic parsing, because each of these forms of analysis expects a monolingual corpus [Piotrowski 2012]. Multilingual tagging may enable separation of the corpus for these monolingual forms of analysis.

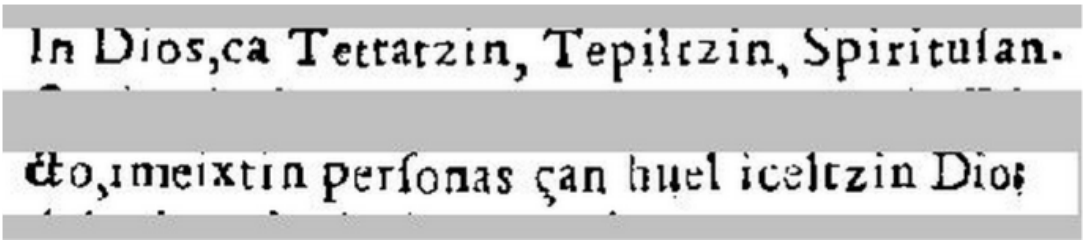


Figure 2. Lines from the *Advertencias* with an automatically produced Ocular transcription. Bold marks Spanish, while underlined is Latin and black is Nahuatl. Note the difference between the two instances of the Spanish word "Dios".In **Dios, cá** Tettatzin, Tepiltzin, Spiritu ſan- cto, in ieixtin perſonas çan huel iceltzin Dios

At the same time, errors in language tagging reveal how these analytic approaches carry their own colonial dangers. Errors in language tagging frequently occur in the Rincón transcription, which often intersperses Nahuatl prefixes, suffixes, and other elements of word use into Spanish descriptions of grammatical patterns. Here the brevity of the word fragments (two or three letters) doesn't provide enough information to trigger a language shift in the system. More interesting for our purposes, however, are errors like those in Figure 2, which shows a fragment from the *Advertencias* that continues the discussion of the holy trinity described above. Here we see that the Spanish word *Dios*, which appears twice in the fragment, has been identified first as Spanish, then as Nahuatl. Elsewhere on the same page, the Latin words *Sancto* and *Sanctisima* were incorrectly tagged as Nahuatl. In both cases, the incorrect tagging is likely triggered by the frequent presence of loanwords in the Nahuatl, Spanish, and Latin training corpus. Though the *Primeros Libros* may be an exaggerated case, early modern writers were almost all multilingual, and early modern writing frequently switches between Latin and the vernacular, using Latin words to emphasize or highlight key terms in much the same way that Spaniards writing in Nahuatl (or their Nahua assistants) drop in terms like *Sancto* (holy) or *Dios* (god). 60

The accurate tagging of loan words poses a particular problem when there is a mismatch between our language data and the text being transcribed, because, as James Lockhart has shown, the use of Spanish loanwords in Nahuatl is period-specific. As with the other examples in this article, however, the concept of a language tagging error obscures a deeper ambiguity in the language itself. If a text written in Nahuatl uses the word *Dios*, is it accurate to describe that as a Spanish word, or would it be more accurate to describe it as a Nahuatl word adopted after conquest? Should we perhaps understand it as the codeswitching of a bilingual writer for whom the boundaries between the two languages were not fixed? The system forces a single linguistic choice where we may in fact be observing the breakdown of standard linguistic categories. [13] 61

Conclusion

According to the assistive pretext of OCR, the primary purpose of automatic transcription is to improve access to digitized texts. OCR accomplishes this by increasing the distribution and legibility of the texts, making them accessible to screen readers and making them available for digital analysis. Implicitly, this process of transcription is also a neutral one: it recognizes characters, rather than reading them; it transcribes them rather than writing. Under the assistive pretext, acts of interpretation disappear. 62

By situating optical character recognition within a longer history of transcription and by conducting an analysis of several systems on which OCR is based, I have suggested that the assistive pretext obscures more complex analytic processes that occur alongside and within the act of transcription. The complex cultural interactions embedded in early colonial transcriptions are a reminder that copies, like computers, are never neutral: they inscribe the scene of their production on the printed page. Similarly, the machine-recognition of printed characters is a historically charged event, in which the system and its data conspire to embed cultural biases in the output, or to affix them as supplementary information hidden behind the screen. 63

A broader implication of this argument is that the machinic transformations of text are neither more nor less radical than those produced by the many systems of interpretive reproduction that came before. Instead, much like the early colonial processes of transcribing and printing the *Primeros Libros*, machines transcribe the historical contingencies of their use into processed texts. This matters for those interested in critical textual studies because it suggests a mode for analyzing digital texts as historical texts. By understanding the historical context of the tools through which digital texts are processed and displayed, we can bring to the surface the traces these tools leave on digital documents. This may allow us to recognize 64

how digital projects implicitly reinforce historical hierarchies of power and knowledge even as our scholarship works to critique or counteract those systems.

This research also matters for those who work to digitize historical documents and seek to produce intellectually rigorous transcriptions. An attentive approach to the processes through which texts are transcribed – from gathering language data to selecting options and transcription tools – will enable practitioners to manage the biases of digitization systems. Rather than being victims of the contingencies of the tools we use, we can target the outcomes we want by considering up front how tools shape text and how historical expectations frame textual reproduction. Building multilingual capabilities into Ocular, for example, allowed us to change the shape of early modern OCR by bringing multilingual documents to the center of automatic transcription rather than imposing an anachronistic, monolingual concept of textuality onto the early modern period. Manually altering our language data similarly allowed the system to be attuned to the unique orthographic qualities of the books in our corpus. 65

Finally, we see a great deal of potential in the amplification of the implicit textual analysis that accompanies OCR. We often treat automatic transcription as a preliminary step to digital analysis, a costly but necessary evil that introduces needless errors without providing much analytic gain. By working with Ocular's developers, we hope to expand Ocular's analytic capabilities, enabling OCR to do more than recognize characters and output transcriptions. Rather than seeking to minimize the potential damage done by OCR's analytic systems, we hope to capitalize on those processes through explicit attention to what machine learning is capable of and where its limitations lie. 66

Acknowledgements

In addition to the many individuals named in this article who have supported this project, I am indebted for their editorial support to Roanne Kantor, Matt Cohen, Kelly McDonough, Dan Garrette, Maria Victoria Fernandez, and Kent Norsworthy. Special thanks to DHQ managing editor Duyen Nguyen and my anonymous reviewers. 67

Notes

[1] Lisa Gitelman's work on the document is a useful reminder of the many kinds of printed text that may exist beyond the narrow standards of the printed book or newspaper [Gitelman 2014].

[2] This paper focuses on Nahuatl, the most heavily represented American language in our corpus. We hope to extend the work for other indigenous languages in future work.

[3] This article follows Laura Mandell's approach to periodization in the digital humanities: "we need to speak of periods still but extending them to take into account, not the reigns of monarchs but dominance of medium" [Mandell 2013, 90]. Here the term "early modern" is used to designate a period across Europe and its colonies during which printed texts shared the defining material qualities of our corpus: wandering baselines, uneven inking, unfamiliar fonts, archaic orthographies, and multilingualism. Mandell uses the dates 1473-1800 to describe this period, with some location-specific flexibility on the upper register.

[4] Since this article was composed, the OCR prototype described here received funding for further development in the form of an NEH Digital Humanities Implementation Grant. The implications of funding structures for Digital Humanities work are beyond the scope of this article, but are worthy of consideration, as Ryan Cordell and others have argued [Cordell 2016]. Any views, findings, conclusions, or recommendations expressed in this article do not necessarily represent those of the National Endowment for the Humanities. The project can be followed at sites.utexas.edu/firstbooks/.

[5] For more about the *Advertencias*, refer to Mark Christensen's *Nahua and Maya Catholicism* [Christensen 2013]; Veronica Murillo Gallegos' article "Obras de personajes novohispanos en las Advertencias" [Murillo Gallegos 2011]; and Louise Burkhart's *The Slippery Earth* [Burkhart 1989]. Christensen is concerned primarily with contextualizing Bautista's ideas of confession within the context of confessional practices in Europe and New Spain. Murillo Gallegos considers the way that the *Advertencias* situates itself within a developing local intellectual culture. And Burkhart examines the decisions that Spanish friars made about how to translate complex theological concepts like purity or sin in order to understand how they navigated or engaged with indigenous cosmologies. She finds that both the preservation of Spanish (or Latin) vocabulary and the adoption of indigenous translations carry ideological weight.

[6] The other two are Andrés de Olmos' *Arte para aprender la lengua mexicana* (1547) and Antonio de Molina's *Arte de la lengua mexicana y castellana* (1571).

[7] This article does not consider other prominent OCR systems, such as Google's Tesseract or ABBYY Fine Reader, though a comparative study could yield interesting results. Both systems do offer the option of a language model, suggesting that some of the implications of this study would be broadly applicable.

[8] "The gerund, endo, is also used in two ways. The first is in the composition of all the verbs that signify stillness or movement, for example, '*ni tetlaçotlatica*, I am loving, *nitetlaçotlatiuitz*. I come loving….'" I have retained the orthographic idiosyncrasies of the original. Thanks to Adam Coon for his help with this translation.

[9] A technical discussion of further modifications to Ocular, made after the composition of this article, can be found in the *Proceedings of NAACL 2016* [Garrette 2016]. This extension enables Ocular to automatically generate a "modernized" text alongside the diplomatic transcription.

[10] It is not mere coincidence that both racialization and colonization are embedded in this title. See Ayhan Aytes, "Return of the Crowds: Mechanical Turk and Neoliberal States of Exception" in *Digital Labor*, ed. Trebor Scholz, 2013; and Shawn Wen, "The Ladies Vanish" in *The New Inquiry* (Nov. 11, 2014).

[11] See Mara Mills' claims about the relationships between reading and blindness; reading and artificial intelligence [Mills 2013] [Mills 2012].

[12] http://www.gutenberg.org/ebooks/11, http://www.gutenberg.org/ebooks/12

[13] It has been noted that this difficulty could be resolved by tweaking the system to allow for an "ambiguous" or "multilingual" tag in cases of language uncertainty. While we will experiment with this in future iterations of the system, broader questions about historical linguistic categories remain relevant.

Works Cited

Angwin et al. 2016 Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. "Machine Bias." *ProPublica (May 23, 2016).*

Aytes 2013 Aytes, Ayhan. "Return of the Crowds: Mechanical Turk and Neoliberal States of Exception." *Digital Labor*. Ed. Trebor Scholz. New York: Routledge, 2013.

Baird 1993 Baird, Ellen T. The drawings of Sahagún's *Primeros memoriales*. Norman: The University of Oklahoma Press, 1993.

Baldridge 2015 Baldridge, Jason. "Machine Learning and Human Bias: An Uneasy Pair." *Techcrunch.com* (August 7,2015).

Bautista 1600 Bautista, fray Juan. *Advertencias. Para los confessores de los Naturales.* Mexico: M. Ocharte, 1600. Primeros Libros. Web. 28 Mar. 2014.

Berg-Kirkpatrick et al. 2013 Berg-Kirkpatrick, Taylor, Greg Durrett, and Dan Klein. "Unsupervised Transcription of Historical Documents." *Proceedings of ACL*. 2013.

Blackburne 1780 Blackburne, Francis. *Remarks on Johnson's life of Milton : to which are added, Milton's Tractate of education and Areopagitica*. London: Hector McLean, 1780. Archive.org. John Adams Library at the Boston Public Library.

Burkhart 1989 Burkhart, Louise. *The Slippery Earth: Nahua-Christian Moral Dialogue in Sixteenth-Century Mexico*. Tucson: University of Arizona Press, 1989.

Burns 2010 Burns, Kathryn. *Into the Archive: Writing and power in colonial Peru*. Durham: Duke University Press, 2010. (4).

Bustamante 1830 Bustamante, Carlos María de. Al editor al que leyere. *Historia general de las cosas de Nueva España*. By Bernardino de Sahagún. Carlos María de Bustamante, Ed. México: Imprenta del Ciudadano Alejandro Valdés, 1829-1830.

Cain Miller 2015 Cain Miller, Claire. "Algorithms and Bias." *The New York Times* (August 10,2015).

Chimalpahin 1889 Chimalpahin Quauhtlehuanitzin, Domingo Francisco de San Antón Muñón. *Annales de Domingo Francisco de San Anton Munon Chimalpahin Quauhtlehuanitzin*. Trans. Siméon Rémi. Paris: Maisonneuve et C. Leclerc, 1889. Archive.org.

Christen 2012 Christen, Kimberley. "Does Information Really Want to be Free? Indigenous Knowledge Systems and the Question of Openness." *International Journal of Communication* 6 (2012): 2870–2893.

Christensen 2013 Christensen, Mark. *Nahua and Maya Catholicisms: Texts and Religion in Colonial Central Mexico and Yucatan*. Palo Alto: Stanford University Press, 2013.

Cordell 2016 Cordell, Ryan. "'Q i-jtb the Raven': Taking Dirty OCR Seriously." Modern Language Association. Austin, Texas. 9 Jan. 2016. ryancordell.com

Eder 2012 Eder, Maciej. "Mind your corpus: systematic errors in authorship attribution." *digital humanities 2012*. 2012.

Ehrman 2012 Ehrman, Bart D. and Michael W. Holmes, eds. *New Testament Tools, Studies and Documents : The Text of the New Testament in Contemporary Research : Essays on the Status Quaestionis. Second Edition (1).* Brill, 2012.

Ernst 2011 Ernst, Wolfgang. "Media Archaeography: Method and Machine versus History and Narrative of Media." *Media Archaeology : Approaches, Applications, and Implications*. Ed. Erkki Huhtamo and Jussi Parikka. Berkeley: University of California Press, 2011. 239–255.

Flanders 2009 Flanders, Julia. "The Productive Unease of 21st-century Digital Scholarship." *Digital Humanities Quarterly* 3.3 (2009).

Garrette 2015 Garrette, Dan, Hannah Alpert-Abrams, Taylor Berg-Kirkpatrick and Dan Klein. "Unsupervised Code-Switching for Multilingual Historical Document Transcription." *Proceedings of NAACL*. 2015.

Garrette 2016 Garrette, Dan and Hannah Alpert-Abrams. "An Unsupervised Model of Orthographic Variation for Historical Document Transcription." *Proceedings of NAACL. 2016.*

Gitelman 2014 Gitelman, Lisa. *Paper Knowledge: Toward a Media History of Documents (Sign, Storage, Transmission).* Durham: Duke University Press, 2014.

Google 2016 "Google reCaptcha." https://www.google.com/recaptcha/intro/index.html

Hayles 2002 Hayles, N. Katherine. *Writing Machines*. Cambridge: The MIT Press, 2002.

Icazbalceta 1858 García Icazbalceta, Joaquín. *Colección de documentos para la historia de México: Tomo Segundo*. México: Libreria de J. M. Andrade, Portal de Agustinos N. 3, 1858.

Icazbalceta 1886 García Icazbalceta, Joaquín. *Bibliografía mexicana del siglo XVI: catálogo razonado de libros impresos en México de 1539 á 1600, con biografías de autores y otras ilustraciones*. Primera Parte. México: Librería de Andrade y Morales, Sucesores,1886.

Jockers 2014 Jockers, Matt. "Text Analysis with R for Students of Literature". Switzerland: Springer International Publishing, 2014. (100).

Kirschenbaum 2012 Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge: The MIT Press, 2012.

Lockhart 1992 Lockhart, James.*The Nahuas After the Conquest*. Stanford: Stanford University Press, 1992. Print.

Lockhart 2001 Lockhart, James. *Nahuatl as Written: Lessons in Older Written Nahuatl, with Copious Examples and Texts*. Stanford: Stanford University Press, 2001.

Lorang 2012 Lorang, Elizabeth and Brian Pytlik Zillig. "Electronic Text Analysis and Nineteenth Century Journals: TokenX and the Richmond Daily Dispatch." *Texas Studies in Literature and Language* 54.3 (2012): 303–323.

Mandell 2013 Mandell, Laura. "Digitizing the Archive: the Necessity of an Early Modern Period." *The Journal for Early Modern Cultural Studies* 13.2 (2013): 83–92.

McDonough 2014 McDonough, Kelly. *The Learned Ones: Nahua Intellectuals in Postconquest Mexico*. Tucson: The University of Arizona Press, 2014.

Mignolo 1995 Mignolo, Walter. *The Darker Side of the Renaissance: literacy, territoriality, and colonization*. Ann Arbor: University of Michigan Press, 1995.

Mills 2012 Mills, Mara. "Other Electronic Books: Print disability and reading machines." *Unbound: Speculations on the Future of the Book* (30 Apr. 2012).

Mills 2013 Mills, Mara. "Blindness and the History of Optical Character Recognition (OCR)." *MLA 2013*. Unpublished.

Murillo Gallegos 2011 Murillo Gallegos, Verónica. "Obras de personajes novohispanos en las *Advertencias para los confesores de los naturales* de fray Juan Bautista de Viseo." *Anuario de Historia de la Iglesia* 20 (2011): 359–371.

O'Neill 2016 O'Neil, Cathy. Weapons of Math Destruction. New York: Crown, 2016.

Piotrowski 2012 Piotrowski, Michael. *Natural Language Processing for Historical Texts*. California: Morgan & Claypool Publishers, 2012.

Rincón 1595 Rincón, Antonio. Arte Mexicana. Mexico: Pedro Balli, 1595. Accessed through the *Primeros Libros* online archive, at www.primeroslibros.org.

Risam and Koh 2012 Risam, Roopika and Adeline Koh. "Postcolonial Digital Humanities: Mission Statement." Web. http://dhpoco.org/mission-statement-postcolonial-digital-humanities/

Robinson 2013 Robinson, Peter. "Towards a Theory of Digital Editions." *Variants* 10 (2013): 105–131.

Schantz 1982 Schantz, Herbert F. *The history of OCR, optical character recognition*. New York: Recognition Technologies Users Association, 1982.

Swift 1823 Swift, Jonathan. *Gulliver's Travels*. Chicago: Hector McLean, 1823. Archive.org. New York Public Library.

Trettien 2013 Trettien, Whitney Anne. "A Deep History of Electronic Textuality: The Case of *English Reprints Jhon Milton Areopagitica*." *Digital Humanities Quarterly* 7.1 (2013).

Tuhiwai-Smith 2012 Tuhiwai-Smith, Linda. *Decolonizing Methodologies*. New York: Zed Books, 2012.

Wakelin 2014 Wakelin, Daniel. *Scribal Correction and Literary Craft: English Manuscripts 1375–1510*. Cambridge: Cambridge University Press, 2014.

Wen 2014 Wen, Shawn. "The Ladies Vanish." *The New Inquiry* (Nov. 11, 2014).http://thenewinquiry.com/essays/the-ladies-vanish/

"Optical Character Recognition" 2015 Wikipedia. "Optical Character Recognition — Wikipedia, The Free Encyclopedia." 2015.